Federal Defenders
OF NEW YORK, INC.

Southern District
52 Duane Street-10th Floor, New York, NY 10007
Tel: (212) 417-8700 Fax: (212) 571-0392

David E. Patton
Executive Director

Southern District of New York
Jennifer L. Brown
Attorney-in-Charge

May 4, 2016

BY ECF
Honorable Valerie E. Caproni
United States District Judge
Southern District of New York
40 Foley Square
New York, New York 10007

Re: United States v. Kevin Johnson
15 Cr. 565 (VEC)

Dear Judge Caproni:

We received the Government's letter dated May 3, 2016, which advised the Court and counsel that the New York City Office of the Chief Medical Examiner (OCME) indicated to the Government that it will not voluntarily disclose the source code for the Forensic Statistical Tool ("FST"). As the Court is aware, we are seeking to exclude evidence generated by FST, or in the alternative, to grant a hearing pursuant to Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 589 (1993). To that end, Rule 17 of the Federal Rules of Criminal Procedure authorizes the issuance of a subpoena for the source code underlying FST.

Statistics of Reporting DNA Results

Reporting the results of a DNA match of a full known profile to a crime scene evidence sample involving a single contributor, or to a mixed sample from which a major contributor can be readily deduced, has for decades involved a statistic known as the Random Match Probability (RMP). Turning on well-settled principles of population genetics, the statistic answers the question of the likelihood of finding the same DNA profile if one were to pick randomly within a given population. The chances are invariably remote.

Interpreting non-deducible mixed samples takes DNA forensics to the edge of its reliability. For mixtures where the major contributor cannot be deduced a relatively recent trend has been to employ a likelihood ratio (LR) to interpret the weight of the evidence. The LR purports to quantify the probability of the evidence under two alternative hypotheses about the source of one of the theoretical contributors to the mixture. The numerator is typically assigned the "prosecutor's hypothesis"(H_p), while the denominator is assigned the "defense hypothesis"(H_d). Set forth most simply, the formula is the probability (Pr) of the evidence (E) given either hypothesis, as here:

$$LR = \frac{\underline{Pr(E|H_p)}}{Pr(E|H_d)}$$

There is no generally accepted, single method for arriving at a LR to evaluate a DNA profile derived from a mixed forensic evidence sample. The reasons for this are manifold. Many relevant variables impact the LR. A non-exhaustive list includes the amount and quality of DNA that is observed in a sample, the number of potential contributors believed to be present in the mixture, the ratio of each contributor's DNA in the sample, the age and quality of the DNA of each contributor in the sample, the relative frequency of the genetic types that are seen in the sample, the quality of the testing that was conducted in the specific case, the method by which stochastic phenomena (allelic drop-out and allelic drop-in) are accounted for by the LR algorithm as well as the specific rates such effects are assigned, and the relative number of specific known and unknown contributors assigned to each of the fundamental propositions H_p and H_d (i.e. for a purported three person mixture, the H_p might be the defendant and two unknown contributors, while the imposed H_d might be three unknown contributors). When applied to a complex, multi-contributor mixed sample comprised of sixteen forensic loci, a multifactorate LR computation becomes intensely demanding. The LR product has been shown to be highly sensitive to influence by changes to even a single variable.[1]

Just how such a highly sensitive, extremely complex calculation should or could be performed is an emergent subject. Because the LR is theoretical, in that it compares two hypotheses, neither of which is inherently subject to objective testing, the principal method for promoting reliable results is open scientific exchange. In the field of complex mathematics, this means open source software. Thus, freely available source code is widely supported in the development of probabilistic genotyping software. For example, the International Society for Forensic Genetics (ISFG), the very same organization cited by the FST's creators to justify the reporting of a LR, supports open source tools, and provides links to open source resources on the front page of its website. Promoting open data exchange: "Open-source is strongly encouraged since this solution offers unrestricted peer review and best assurance that methods are fit for purpose."[2] In other words, "[w]e do not advocate a black box approach."[3]

[1] See e.g. P. Gill, et al, DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures. FORENSIC SCI. INT'L: GENETICS 2006; July 13, 2006, 160, at 90-101.

[2] P. Gill, et al, DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods, FORENSIC SCI. INT'L: GENETICS 6, June 3, 2012, at 684.

[3] Id.

OCME's Application of FST to Likelihood Ratio Reporting

FST is a closed source, "black box" software program that generates a LR for mixed samples where the major contributor cannot be deduced.[4] Comparing a known profile to a mixed sample, the program generates a LR to describe the weight of the evidence. Published studies regarding FST do not reveal the formulae by which FST reaches the LR. What is known is that FST employs fixed variables for drop-out and drop-in rates. What is not known is how those rates interact with other variables in the software, or whether they are updated with new information from the years of testing the OCME has conducted since the validation of FST was completed. The method by which FST calculates drop-in also appears from published literature to both under-estimate the likelihood that drop-in has occurred, and under-represent the number of potential contributors to the mixture.[5]

The template amount of DNA in a given sample is an apparently critical variable to the LR result, yet FST was validated against only six sample sizes: 25pg, 50pg, 100pg, 150pg, 250pg, and 500pg. None of the samples in this case fit one of these size standards. When a sample falls between a standard range, FST "interpolates." How this affects the LR has not been published or subject to review and is not understood.

Additionally, the way FST incorporates population genetics cannot be subject to meaningful review given the current state of production. The program "adjusts only for intra-individual correlation with a correction to the expected population frequencies of homozygous

[4] The OCME's proprietary stance as to FST's code not only contrasts with ISFG policy, it is bad science in this developing forensic application and thus seems inconsistent with the mandate of a public lab – especially in the criminal justice context where it is entirely predictable that reported results will be subject to the discovery process and further expert inquiry, testing, and scrutiny are a routine matter. "Our view is that we have reached the point that, with some exceptions, anything less than release of actual source code is an indefensible approach for any scientific results that depend on computation, because not releasing such code raises needless, and needlessly confusing, roadblocks to reproducibility." Darrel C. Ince et al. The Case for Open Computer Programs, NATURE, Feb. 23, 2012, at 485 (available at http://www.nature.com/nature/journal/v482/n7386/pdf/nature10836.pdf). The authors, in arguing for the release of source code underlying scientific findings, argue that in the absence of code, at least one several problems is likely to occur: vague, ambiguous descriptions, programming errors, errors associated with numerical properties of scientific software, well-known ambiguities in some internationally standardized versions of commonly used programming languages in scientific computation, and problems with machine deployment. Id. at 485-7.

[5] Adele Mitchell, et.al, Validation of a DNA mixture statistics tool incorporating allelic drop-out and drop-in. "In addition, LoComatioN and LRMix model allelic drop-in using estimated allele frequencies, whereas FST does not consider the identity of drop-in alleles, simply that drop-in of one allele or of two or more alleles has occurred." FORENSIC SCI. INT'L: GENETICS 6. August 9, 2012, at 749-750.

genotypes."[6] This statement may have a discernable mathematical meaning, but it is not readily apparent. There are known definitions to the vocabulary employed, and perhaps an intended meaning in the published literature, but it is not sufficiently clear to comprehend how FST actually functions. For example, what is a "correction" in this context? How much of a correction is adjusted, and what effect does that have on the LR? This cannot be drawn from the published literature, discovery, or validation. Like each other dependent variable touched upon above, the answer is embedded within FST's source code.

The Legal Standard

Rule 17 authorizes the production by a witness of "books, papers, documents, data, or other objects the subpoena designates." Fed. R. Crim. P. 17(c). In interpreting this rule, courts have evaluated the (a) reliability, (b) admissibility, and (c) specificity of the target of the subpoena. United States v. Nixon, 418 U.S. 683, 700 (1974). The moving party is required to show: (1) that the documents are evidentiary and relevant; (2) that they are not otherwise procurable reasonably in advance of trial by exercise of due diligence; (3) that the party cannot properly prepare for trial without such production and inspection in advance of trial and that the failure to obtain such inspection may tend unreasonably to delay the trial; and (4) that the application is made in good faith and is not intended as a general fishing expedition. Id. at 699-700 (relying on United States v. Iozia, 13 F.R.D. 335, 338 (S.D.N.Y. 1952)). Rule 17 also dictates that a court may "quash or modify a subpoena if compliance would be unreasonable or oppressive." Fed. R. Crim. P. 17(c)(2).

Here, the FST source code satisfies the requirements of Nixon and any of the less restrictive standards and therefore, a Rule 17 subpoena is appropriate. See, e.g., United States v. Ocasio, No. EP-11-CR-2728-KC, 2013 WL 2458617, at *1 (W.D. Tex. June 6, 2013) (granting motion to compel and denying motion to quash a request for subpoenas seeking the source code from a third party software company when defendant challenged the government's use of 'child protection system' software in a Fourth Amendment suppression motion).

Such a subpoena is particularly necessary given the challenge at hand: a motion *in limine* pursuant to Federal Rule of Evidence 702 and Daubert. Both Rule 702 and Daubert demand that the Court serve a "gatekeeping" function. Thus, before a FST-generated likelihood ratio can be introduced at trial, the Court must be satisfied that it clears the bars set forth in Rule 702: expert testimony is permitted if it is (1) is based upon sufficient facts or data, (2) it is the product of reliable principles and methods, and (3) the expert testifying has applied the principles and methods reliably to the facts of the case, and set forth in Daubert: the Court must "ensure that any and all scientific testimony or evidence admitted is not only relevant, but reliable." Daubert, 509 U.S. at 589. As part of its gatekeeping function, the district court must make a "preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid and of whether that reasoning or methodology properly can be applied to the facts in issue." Daubert, 509 U.S. at 592-93.

[6] Id. (comparing FST with probabilistic software that also incorporates adjustments for inter-individual correlation of genotypes).

In order for the Court to make such a determination in the instant case, we intend to offer the Court a comprehensive challenge to the reliability of FST. The only meaningful way to do that is to have our experts evaluate the software. Because the software is "black box" or "closed source," the results are not reproducible – i.e. they cannot be tested for accuracy and reliability. Accuracy and reliability are the lynchpins of our 702 and Daubert challenge.

As discussed, a slight variability in any parameter in the numerator or denominator will dramatically impact the LR. We seek to understand how the FST software employs imbedded variables, for example, the drop-out and drop-in rates, how FST "interpolates" the template amount of the mixed samples in Mr. Johnson's case in its unique path to the LR. Without access to the source code, the computations done by FST are unknowable – and if they are unknowable, they cannot be evaluated under the 702 and Daubert standards. As the Government points out, in the face of the inscrutability of FST, some defense attorneys in state court have gone so far as to attempt to reverse engineer FST, with differing results. This desperate effort only underscores the necessity of our request. Only with the actual program code can the parties meaningfully brief, and can the Court fully evaluate, the Daubert factors, which can include, but are not limited to: (1) whether the theory or technique "can be (and has been) tested;" (2) whether the theory or technique "has been subjected to peer review and publication;" (3) a technique's "known or potential rate of error;" and the "existence and maintenance of standards controlling the technique's operation;" and (4) whether a particular technique or theory has gained "general acceptance" in the "relevant scientific community." Daubert, 509 U.S. at 593-94. Those factors and others are squarely addressed by an evaluation of the source code. As New York State does not follow a Daubert standard, the state court decisions cited by the Government are inapposite.

The fact that other relevant material has been produced in discovery is of no import to this request. In fact, it is precisely from our intensive review of those materials that we concluded that access to the source code would be necessary. The value of the validation studies, for example, is minimal without connecting those materials to the program itself. Moreover, the available operating procedures and user manuals only speak in general terms to how FST is supposed to work, but do not address the central question here: does the software reliably do what it purports to do?

The OCME further objects on the grounds that FST is a "proprietary and copyrighted" tool. This objection simply is of no moment, as any property or copyright interests can be fully protected by an appropriate protective order. Your Honor has already indicated that, should the subpoena we request be issued, such an order would be a condition. Courts readily exercise this discretion when necessary to ensure both that sensitive information is protected and that the defendant's right to present a full defense is preserved. See, e.g., United States v. Diakhoumpa, No. 15-CR-629, 2016 WL 1105486, at *3 (S.D.N.Y. Mar. 15, 2016) (ordering government to produce under protective order trade secret information provided by trademark holders to government's experts in smuggling and counterfeiting case); United States v. Durst, Crim. No. 15-091, 2015 WL 4879465, at *3-4 (E.D. La. Aug. 14, 2015) (denying motion to quash subpoena seeking proprietary trade secrets and issuing protective order to protect contents from public dissemination). Even clearly commercial interests, which we are here left to infer from the OCME's assertion of copyright protections, can be sufficiently protected by an appropriate order. See United States v. Siegel, No. 96 CR. 411, 1997 WL 12804, at *4 (S.D.N.Y. Jan. 14, 1997) (in a criminal antitrust action, where disclosure of 3rd party competitor's bid formula would give defendant significant commercial advantage in the future, the court granted production only to counsel and an appropriate expert.)

In the context of civil litigation, courts regularly order the production of valuable trade secrets, specifically including source code, subject to orders that protect the disclosing party from commercial disadvantage. See, e.g., Dynamic Microprocessor Assoc. v. EKD Computer Sales, 919 F. Supp. 101, 106 (E.D.N.Y. 2007) (ordering plaintiff to produce source code in copyright infringement litigation but limiting availability of source code to defense counsel and defendant's expert); Quotron Sys., Inc. v. Automatic Data Processing, Inc., 141 F.R.D. 37, 41 (S.D.N.Y. 1992) (denying motion to quash subpoena seeking programming information involving "trade secrets between competitors" noting that "confidentiality concerns raised by these subpoenas can be dealt with by means of a protective order."); see also Brown Bag Software v. Symantec Corp., 960 F.2d 1465 (9th Cir. 1992) (affirming scope of protective order fashioned by district court to protect trade secrets, including source code).[7] Indeed, some of the world's most valuable trade secrets and intellectual property have been produced under appropriate protective orders. See, e.g., Coca-Cola Bottling Co., of Shreveport, Inc. v. Coca-Cola Co., 107 F.R.D. 288, 289, 300 (D. Del. 1985) (ordering disclosure of Coca-Cola's secret formula – "one of the best-kept trade secrets in the world" – under a protective order that "both allows access to information and prevents disclosure of trade secrets"). As with Apple, Google, and Coca-Cola, what property interests the OCME has in FST can be readily protected by an appropriate protective order.

The Government mischaracterizes our request as a "fishing expedition", relying on a petty offense case, United States v. German, No. 4:12-mj-275, 2013 WL 5347360, at *1-2 (M.D. Ga. Sept. 23, 2013). There, however, defense counsel filed a two page motion to compel, and thereafter failed to timely object to the proprietor of the source code at issue's proposed protective order. Even after the court denied the defendant's initial motion, the proprietor of the code in fact made the code available to the defendant, on the condition that he agree to a protective order. The defendant chose to decline that offer. It was only after the defendant chose not to enter into that protective order, and again sought production of the source code - this a mere ten days prior to trial - that the Court deemed the late-breaking request a "fishing expedition" and granted the motion to quash. The instant request is readily distinguishable from German, principally in that we are seeking precisely what the proprietor in that case agreed to provide.

[7] Parties also routinely stipulate to protective orders governing highly sensitive source code. See, e.g., Stipulated Protective Order Regarding Disclosure and Use of Discovery Materials, Jongerius Panoramic Technologies, LLC v. Google Inc., No. 12-03797, Feb. 13, 2013, ECF No. 137 (ordering stipulation between Google, Apple and Plaintiff to govern confidential information, including provisions governing source code). Indeed, some District Courts' model protective orders include such provisions. *See* United States District Court for the Northern District of California, Model Stipulated Protective Order for Litigation Involving Patents, Highly Sensitive Confidential Information and/or Trade Secrets, available at http://www.cand.uscourts.gov/filelibrary/776/ND_Cal_Patent_Highly_Sensitive_Model_Prot_Ord_Revised.docx.

Finally, the Government's letter indicates that the source code is "over 500 pages." Such a production – which is the approximately one hundred pages less than the United States Sentencing Commission Guidelines Manual – cannot reasonably be claimed to be burdensome.

For these reasons, our request for a subpoena for the FST software should be granted.

Respectfully submitted,

Christopher Flood

SZ

Sylvie Levine
Assistant Federal Defenders